SANDIA REPORT

SAND98–1769/1
Unlimited Release
Printed August 1998

Statistical Considerations in Designing Tests of Mine Detection Systems: I – Measures Related to the Probability of Detection

Katherine M. Simonson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Prices available from (615) 576-8401, FTS 626-8401

Available to the public from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

SAND98-1769/1
Unlimited Release
Printed August 1998

Statistical Considerations in Designing Tests of Mine Detection Systems: I - Measures Related to the Probability of Detection

Katherine M. Simonson
Signal & Image Processing Systems Department
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0844

Abstract

One of the primary metrics used to gauge the performance of mine detection systems is PD, the probability of detecting an actual mine that is encountered by the sensor. In this report, statistical issues and techniques that are relevant to the estimation of PD are discussed. Appropriate methods are presented for summarizing the performance of a single detection system, for comparing different systems, and for determining the sample size (number of target mines) required for a desired degree of precision. References are provided to pertinent sources within the extensive literature on statistical estimation and experimental design. A companion report addresses the estimation of detection system false alarm rates.

This work performed under sponsorship of the DoD/DOE
Memorandum of Understanding on Non-Nuclear Munitions Technology

CONTENTS

LIST OF FIGURES

1 - INTRODUCTION

A number of statistical concepts are directly relevant to the design of mine detection experiments and demonstrations, and to the analysis of the data collected during such exercises. In particular, standard statistical calculations can be used to bound the uncertainty that will be present in system performance measures derived from experimental data. For mine detection systems, the primary metrics used to gauge performance are related to PD (the probability of detecting actual mine targets), and FAR (the false alarm rate). Of course, other features (e.g., cost, speed, and operator safety) are also important, but these are largely non-statistical concerns. In this paper, statistical models and computations applicable to the measurement of PD are outlined. An appropriate statistical framework for assessing FAR is discussed in a companion report [Simonson, 1998].

Typically, two types of statistical questions are of interest when analyzing performance data. These are questions of *estimation* and questions of *comparison*. An example of an estimation question is: "What is the probability that detection system A will find a particular type of target?" An example of a comparison question is: "Does system A have a different PD than system B under specified conditions?" Experiments can be designed to answer either of these types of questions to a desired level of precision. When (as is usually the case) the size of the experiment is constrained by cost, one can design to ensure adequate precision for the most important estimates and comparisons, while addressing less critical questions with lower precision.

In the next section, the binomial model for data related to PD is introduced. Issues and calculations related to estimation and testing for a single binomial parameter are presented in section 3. Methods of constructing confidence intervals, conducting hypothesis tests, and selecting sample sizes are described, and illustrative examples are included. Section 4 covers topics related to questions of paired comparison (one system versus another) under the binomial model. A few caveats and general considerations are briefly discussed in Section 5.

2 – THE BINOMIAL MODEL

2.1 - Background and Assumptions

The binomial distribution [Johnson, Kotz, and Kemp, 1992] is generally used to model experimental data related to the probability of detecting actual mine targets. Each binomial "trial" represents the presentation of a single mine to the detection system, and a full "experiment" consists of a number of different trials. At each trial, two outcomes are possible: either the mine is detected, or it is missed. The experimenter must develop a reasonable protocol for determining specifically what constitutes a detection.

The binomial model assumes that a detection system being tested has a fixed (but unknown) probability of detecting a particular type of mine under specified conditions. In statistics, this probability is usually represented by the parameter *P*; it corresponds to the quantity PD in target recognition terminology. The individual trials are assumed to be identically distributed, and independent from one another.

These latter assumptions merit some discussion. In order to ensure that all of the trials used to estimate PD under given circumstances are roughly identically distributed, efforts should be made to control any variables thought to influence performance. For example, individual mines should be similar in size and physical properties, burial depth should be about the same from mine to mine, and soil chemistry should not change greatly. Some variation in these characteristics will always remain, but this should be minimized to ensure applicability of the model. When substantial sources of variation exist (e.g., low metallic versus nonmetallic mines, clay versus sand), the performance analysis should be carried out separately for each set of factors. Statistical tests can then be conducted to determine which of the experimental factors has a significant effect on performance.

The assumption that individual trials are independent is generally a reasonable one, provided that each target mine is encountered but once, and mines are placed far enough apart to prevent signal interference. If the full experiment is to be replicated, so that detectors encounter

the same buried mine two or more times, one should test for a "learning" effect before pooling the data from multiple runs. Replication can provide valuable information about performance consistency, but appropriate cautionary measures must be taken both in conducting the experiment [Andrews, George, and Altshuler, 1997], and in analyzing the resulting data.

2.2 - Notation

In a full experiment, some number, *n*, of target mines is presented to a mine detection system. Let *X* be a random variable representing the number of these mines that the system detects; *X* is said to have the binomial distribution with parameters *n* and *P* [Johnson, Kotz, and Kemp, 1992]. One major goal of the exercise is to make inferences about the parameter *P*.

The percentage of mines detected is given by:

$$\hat{P} = \frac{X}{n}. \tag{1}$$

The quantity $\hat{P}$ is an estimator of the true probability *P*. Intuitively, the uncertainty inherent in this estimator will decrease as *n* increases: an experiment giving 50 out of 75 detections is far more informative than an experiment giving 2 out of 3. The mean and variance of $\hat{P}$ are as follows:

$$E\left(\hat{P}\right) = P \tag{2}$$

$$\mathrm{var}\left(\hat{P}\right) = \frac{P(1-P)}{n} \tag{3}$$

The variance of an estimator, defined as the expected squared deviation of a value about its mean, is a measure of the uncertainty present in that estimator. The denominator of (3) reinforces the intuitive notion that uncertainty decreases as the sample size *n* increases. Consistency also affects variance. For fixed *n*, the variance is maximized when *P* = 0.50. It is

minimized (and equals zero) when the outcome of each trial is certain, at $P = 0.0$ and $P = 1.0$. These observations are prominently employed both in experimental design and in data analysis.

3 – ESTIMATION AND TESTING FOR A SINGLE BINOMIAL PARAMETER

3.1 - Confidence Intervals for *P*

The use of confidence intervals is one standard way to report parameter estimates along with uncertainty measures. Confidence intervals are computed from observed data, and thus are constructed after completion of the experiment. Informally speaking, a confidence interval contains the values of the unknown parameter that are consistent with the observed data. Consistency is measured in terms of the a priori probability that the computed interval will contain the true value of the parameter: 95% confidence intervals are constructed in such a manner that they will have a 95% chance of containing the true value. More formal definitions of confidence intervals abound [Bickel and Doksum, 1977; Cox and Hinkley, 1974; Silvey, 1975].

The degree of confidence in an interval is often represented algebraically in terms of the quantity α, which is equal to one minus the a priori probability that the interval will contain the true parameter value. Thus, for a 95% confidence interval, α is equal to 0.05. The standard notational convention uses the expression $100(1 - \alpha)\%$ to represent the certainty corresponding to a generic confidence interval.

Two different approaches are employed to construct confidence intervals for a single binomial parameter, P. When the number n of trials is large, the (continuous) normal distribution is used to approximate the (discrete) binomial [Fleiss, 1981; Lehmann, 1975]. This approximation makes the construction of confidence intervals straightforward for large n. However, if n is small the normal approximation is inappropriate and a more cumbersome format for confidence intervals is required. Various authors [Fleiss, 1981; Hahn and Shapiro, 1967] suggest that the large-sample approach is acceptable when both nP and $n(1 - P)$ exceed five.

The article by Blyth and Still [1983] provides a thorough discussion of the major forms of binomial confidence intervals, including the versions recommended here.

3.1.1 - Small Sample Approach

The small sample method for computing confidence intervals for the binomial parameter is as follows. Denote the lower limit of an interval by P_L, and denote the upper limit by P_U. If n trials are conducted and x mines are detected, an approximate $100(1 - \alpha)\%$ confidence interval for P is defined by:

$$P_L = \frac{\nu_1 F_{\nu_1,\nu_2,\alpha/2}}{\nu_2 + \nu_1 F_{\nu_1,\nu_2,\alpha/2}} \tag{4}$$

$$P_U = \frac{\nu_3 F_{\nu_3,\nu_4,1-\alpha/2}}{\nu_4 + \nu_3 F_{\nu_3,\nu_4,1-\alpha/2}}\ , \tag{5}$$

with $\nu_1 = 2x$, $\nu_2 = 2(n - x + 1)$, $\nu_3 = 2(x + 1)$, and $\nu_4 = 2(n - x)$ [Johnson, Kotz, and Kemp, 1992]. Here, $F_{\nu_1,\nu_2,\alpha/2}$ is equal to the $\alpha/2$ quantile of the F distribution with υ_1 and υ_2 degrees of freedom. Similarly, the quantity $F_{\nu_3,\nu_4,1-\alpha/2}$ is defined as the $1 - \alpha/2$ quantile of the F distribution with υ_3 and υ_4 degrees of freedom. Quantiles of the F distribution are tabulated in many statistics textbooks [e.g. Larsen and Marx, 1981], and are readily available from most commercial statistical software packages.

The formulas (4) and (5) work for $0 < x < n$. When $x = 0$, the bounds:

$$P_L = 0.0 \tag{6}$$

$$P_U = 1 - \alpha^{1/n} \tag{7}$$

are recommended, and when $x = n$ the bounds:

$$P_L = \alpha^{1/n} \tag{8}$$

$$P_U = 1.0 \tag{9}$$

are recommended [Johnson, Kotz, and Kemp, 1992]. Both of these represent 100(1 – α)% confidence intervals for *P*.

A simple example illustrates the small sample method. Suppose that n = 10 trials are conducted, and x = 8 of the mines are detected. The parameters of the relevant F distributions are ν_1 = 16, ν_2 = 6, ν_3 = 18, and ν_4 = 4. In order to construct 95% confidence intervals, we choose α = 0.05. From a table of the F distribution, $F_{16,6,.025}$ = 0.299, and $F_{18,4,.975}$ = 8.592. It follows from (4) and (5) that an approximate 95% confidence interval for the true probability of detection is given by (0.444, 0.975). Values of *P* falling within this interval are deemed to be consistent with the observed data.

Confidence interval width ($P_U - P_L$) is a natural measure of the uncertainty present in an estimate. Figure 1 shows lower and upper 95% confidence bounds, as well as interval widths, for each possible outcome of experiments with n = 10 and n = 20 mines. All of the values plotted are calculated from equations (4) - (9). Note that even for a fixed sample size, width varies considerably as a function of $\hat{p} = x/n$.

3.1.2 - Large Sample Approach

The large-sample method of computing confidence intervals uses the normal approximation to the binomial distribution. For values of $\hat{p}$ close to the endpoints, the preferred form [Fleiss, 1981; Koopmans, 1987] for approximate 100(1 – α)% confidence intervals has endpoints given by:

$$P_L = \frac{(2n\hat{p} + z^2_{1-\alpha/2} - 1) - z_{1-\alpha/2}\sqrt{z^2_{1-\alpha/2} - (2 + 1/n) + 4\hat{p}(n - n\hat{p} + 1)}}{2(n + z^2_{1-\alpha/2})} \tag{10}$$

$$P_U = \frac{(2n\hat{p} + z^2_{1-\alpha/2} + 1) + z_{1-\alpha/2}\sqrt{z^2_{1-\alpha/2} + (2 - 1/n) + 4\hat{p}(n - n\hat{p} - 1)}}{2(n + z^2_{1-\alpha/2})} \quad . \tag{11}$$

Here, the quantity $z_{1-\alpha/2}$ represents the quantile of the standard normal distribution corresponding to probability $1 - \alpha/2$. For example, to get a 95% confidence interval, one would choose $\alpha = 0.05$, and use the value $z_{0.975} = 1.960$ in equations (10) and (11). Tables of the standard normal distribution are found in many statistics textbooks [e.g. Larsen and Marx, 1981], and are readily available from most commercial statistical software packages.

For $\hat{p}$ away from the endpoints (say, $0.30 \le \hat{p} \le 0.70$) a simpler formula for approximate $100(1 - \alpha)$% confidence intervals is appropriate. Here, the bounds are given by:

$$P_L = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \frac{1}{2n} \tag{12}$$

$$P_U = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{2n} \quad . \tag{13}$$

The terms in $1/(2n)$ represent a correction for the continuity of the normal distribution, and are omitted by some authors. Intervals of the general form (12) - (13), with or without the continuity correction, are provided in many references [Fleiss, 1981; Koopmans, 1987; Larsen and Marx, 1981].

Figure 2 shows upper and lower 95% confidence bounds, along with interval widths, for each possible outcome of experiments with $n = 50$ and $n = 100$ target mines. All of the values shown are computed using Equations (10) through (13). As in the small sample case (Figure 1), the confidence interval widths computed here vary with both n and $\hat{p}$.

3.2 - Hypothesis Tests

Statistical hypothesis tests provide a mechanism for choosing among two conflicting hypotheses about the model underlying an observed data set [Koopmans, 1987; Silvey, 1975]. The first hypothesis is referred to as the "null hypothesis" (H_0), and is accepted in the absence of convincing evidence to the contrary. The "alternative hypothesis" (H_1) is accepted when experimental data are deemed to be inconsistent with H_0.

Two types of errors are possible when using an hypothesis test to decide between H_0 and H_1: one may reject H_0 when H_0 is true, or one may accept H_0 when H_1 is true. The "level" of a test, commonly denoted α, is defined to be the probability of making the first type of error. The "power" of a test, denoted $1 - \beta$, is defined to be one minus the probability of making the second type of error. Ideally, one would like to have a test with level $\alpha = 0.0$ and power $1 - \beta = 1.0$. In practice this is rarely possible, so a standard approach is to choose a test with a specified level ($\alpha = 0.05$ and $\alpha = 0.10$ are both common choices), and as much power as possible. The decision to accept or reject H_0 is made based on the value of a "test statistic" computed from the observed data.

In the case of a single binomial parameter, one may wish to test whether observed data are consistent with the hypothesis that P is equal to some specified value, P_0. The appropriate null and alternative hypotheses are given by:

$$H_0 : P = P_0 \tag{14a}$$

$$H_1 : P \neq P_0 \tag{14b}$$

and the following test statistic is used:

$$z = \frac{\left| \hat{p} - P_0 \right| - 1/(2n)}{\sqrt{\frac{P_0(1-P_0)}{n}}}. \tag{15}$$

When n is large and the null hypothesis is true, the statistic (15) has a distribution that is approximately standard normal [Fleiss, 1981]. If H_1 is true, z will tend to be large. An α-level test rejects H_0 when z exceeds $z_{1-\alpha/2}$, the $1 - \alpha/2$ quantile of the standard normal distribution.

As an example, consider a test of H_0: $P = 0.5$ versus H_1: $P \neq 0.5$, and suppose that the available data show $x = 24$ detections in $n = 40$ trials. From (15), the observed value of the test statistic is $z = 1.107$. To test at level $\alpha = 0.05$, compare z to $z_{0.975} = 1.960$. Because $z < 1.960$, H_0 is accepted at level 0.05.

The relationship between α-level hypothesis tests and $100(1 - \alpha)\%$ confidence intervals is often straightforward. The hypothesis test outlined in (14) and (15) directly corresponds to the large-sample confidence intervals defined in (12) and (13). If a detection rate of $\hat{p}$ is observed, then (14a) will be accepted at level α for all values of P_0 lying within the $100(1 - \alpha)\%$ confidence interval, and will be rejected for all values of P_0 lying outside of the interval [Fleiss, 1981]. A number of authors [Bickel and Doksum, 1977; Fleiss, 1981; Koopmans, 1987] probe more deeply into the duality between hypothesis tests and confidence intervals.

Note that the test statistic (15), which is based on the normal approximation to the binomial distribution, is only appropriate when the number of trials is large. For small n, the recommended testing procedure would be to reject (14a) for values of P_0 not lying in the small-sample confidence interval computed from (4) through (9).

3.3 - Sample Size Calculations

The normal approximation to the binomial distribution can be used to calculate the number of trials (mines) needed in a planned experiment, when the uncertainty in estimates of P must be kept below some specified level [Feller, 1950; Koopmans, 1987; Larsen and Marx, 1981]. Here, uncertainty is expressed in terms of $100(1 - \alpha)\%$ confidence interval width. The approach taken is to specify a tolerable width for confidence intervals on P, and then solve for the sample size that will provide this width. This is roughly equivalent to requiring that the

estimate $\hat{p}$ should have a $100(1-\alpha)\%$ chance of lying within $W/2$ of the true (but unknown) value P.

The width of the approximate interval specified by (12) and (13) is given by:

$$W = 2z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{n}. \tag{16}$$

Recall that $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. To compute the required sample size as a function of $\hat{p}$, one can set the width to the desired value and solve (14) for n. Omitting the (relatively inconsequential) $1/n$ term, we find that:

$$n = \frac{4z_{1-\alpha/2}^2\,\hat{p}\,(1-\hat{p})}{W^2}\,. \tag{17}$$

Some complication is introduced by the fact that one does not know the value of $\hat{p}$ until the completion of the experiment. One tactic for resolving this difficulty is to set $\hat{p} = 0.50$, which maximizes (17) over all possible $\hat{p}$. This conservative approach gives a sample size of:

$$n = \frac{z_{1-\alpha/2}^2}{W^2}, \tag{18}$$

and ensures that the confidence interval for P will have a width of W or smaller, for all possible outcomes. For example, suppose that 95% confidence intervals having width no larger than 0.20 are desired for all $\hat{p}$. Since $z_{0.975}$ is equal to 1.960, (17) implies that at least 96 mines must be emplaced.

An alternative to the use of $\hat{p} = 0.50$ in (17) is to choose a value of $\hat{p}$ that is believed to be reasonable based on physical principles or prior testing. Thus, if a particular test scenario is expected to be very "easy" (in the sense of detection probabilities close to unity) or very

"difficult" (detection probabilities close to zero), this scenario will require fewer trials than another scenario for which more variable performance is anticipated. As an example, if the probability of detection is expected to be at least 85% for a particular type of mine, setting $n = 49$ trials should provide a 95% confidence interval with width no larger than 0.20. This reduces by about half the number of trials found using (18).

A final alternative to (18) is applicable if detectors performing below a particular level are of marginal interest. In this case, one may wish to constrain the confidence interval width only for detection systems with performance above some threshold, $\hat{p}^*$. Assuming that $\hat{p}^*$ exceeds 0.50, one can substitute it for $\hat{p}$ in (17). Suppose that one wishes to focus only on systems detecting at least 75% of the target mines. In order to ensure 95% confidence interval widths of 0.20 or less for such systems, at least 72 mines are required. Confidence intervals can still be computed for detection systems with $\hat{p} < \hat{p}^*$, but such intervals may have widths exceeding the desired maximum.

Figure 3 shows the approximate minimum number of trials required to achieve specified 90% and 95% confidence interval widths, as a function of $\hat{p}$. Recall that the limits (12) and (13) are not strictly appropriate when either nP or $n(1 - P)$ is less than five, or when P is close to the endpoints 0 and 1. It follows that minimum sample sizes computed using (17) may not be exact in such cases. This is of minor concern in most instances.

4 – COMPARING TWO BINOMIAL PROPORTIONS

This section presents statistical methods appropriate for use in comparing the experimental results obtained for two different detectors, referred to as system A and system B. Define n_A to be the number of trials conducted for system A, and let x_A be the number of detections. The observed proportion of target mines detected is then $\hat{p}_A = x_A/n_A$. The quantities n_B, x_B, and $\hat{p}_B$ are similarly defined for system B. The true probability of detection for system A is denoted P_A, and P_B is the true probability of detection for system B. Inferences about

comparative performance are based on the difference $P_B - P_A$. Confidence intervals, hypothesis tests, and sample size calculations are all discussed.

4.1 - Confidence Intervals for $P_A - P_B$

As in the case of a single proportion, confidence intervals for the difference between two proportions are not computed in the same manner for large and small samples. The large sample method is appropriate when the sample sizes n_A and n_B are large in the sense that the quantities $n_A P_A$, $n_A(1 - P_A)$, $n_B P_B$, and $n_B(1 - P_B)$ all exceed five [Fleiss, 1981].

4.1.1 - Small Sample Approach

The problem of constructing small sample confidence intervals for the difference between two binomial proportions has received considerable attention in the statistical literature [Beal, 1987; Mantel, 1988; Peksun, 1993; Santner and Snell, 1980] and remains an area of active interest [Wallenstein, 1997]. Unfortunately, the methods proposed are generally quite complex and awkward to implement. While some solutions are available in commercial software [StatXact-3, 1995] there is no general agreement in the literature as to which approach is preferable. No specific recommendation is made here.

4.1.2 - Large Sample Approach

The large sample approach is based on the normal approximation to the binomial distribution, and computation of the confidence intervals is straightforward. A $100(1 - \alpha)\%$ confidence interval for $P_B - P_A$ is defined by the bounds [Fleiss, 1981]:

$$P_L = (\hat{p}_B - \hat{p}_A) - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}} - \frac{1}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \tag{19}$$

$$P_U = (\hat{p}_B - \hat{p}_A) + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}} + \frac{1}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right). \quad (20)$$

As before, $z_{1-\alpha/2}$ represents the quantile of the standard normal distribution corresponding to probability $1 - \alpha/2$.

As an example of the use of (19) and (20), suppose that system A was presented with n_A = 37 target mines, of which x_A = 15 were detected, while system B encountered n_B = 40 mines and detected x_B = 31. It follows that $\hat{p}_A$ = 0.405, and $\hat{p}_B$ = 0.775. To create a 95% confidence interval for $P_B - P_A$, α is set to 0.05. The relevant normal quantile is $z_{0.975}$, which equals 1.960. It follows from (19) and (20) that (0.139, 0.600) represents a 95% confidence interval for the difference in detection probabilities. Note that the value 0.0 (implying equal capability for the two systems) lies outside of the interval.

4.2 - Hypothesis Tests

The performances of two different mine detection systems can be compared in a straightforward fashion by asking the question: "Is the PD demonstrated by system A significantly different from the PD demonstrated by system B?" In statistical terms, this question is phrased as an hypothesis test, with null and alternative hypotheses given by:

$$H_0 : P_A = P_B \quad (21a)$$

$$H_1 : P_A \neq P_B \quad (21b)$$

A standard test statistic for employment in this situation is:

$$z = \frac{|\hat{p}_B - \hat{p}_A|}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}, \quad (22)$$

where $\bar{p} = (n_A \hat{p}_A + n_B \hat{p}_B)/(n_A + n_B)$ [Larsen and Marx, 1981]. The observed value of z is compared to the $1 - \alpha/2$ quantile of the standard normal distribution to determine whether the difference is significant at level α. If z exceeds $z_{1-\alpha/2}$ the null hypothesis is rejected. Otherwise, H_0 is accepted.

The test given by (21) and (22) is equivalent to the familiar Pearson chi-square test [Bickel and Doksum, 1977; D'Agostino, Chase, and Belanger, 1988]. Some authors [Fleiss, 1981; Lehmann, 1975] recommend a variation of the test statistic that incorporates a continuity correction. However, simulation studies [D'Agostino, Chase, and Belanger, 1988] indicate that the uncorrected version is preferable, and is appropriate for both large and small samples.

The method can be illustrated using the hypothetical performance data given in section 4.1.2. For these data, the test statistic takes on value z = 3.304. To test at the 95% level, z is compared to $z_{0.975}$ = 1.960. Because the observed value of the test statistic exceeds the threshold, H_0 is rejected, and the performance difference between the two systems is deemed to be significant at the 5% level.

4.3 - Sample Size Calculations

The test statistic (22) can be used to determine the sample size required to compare the performances of two different systems to a specified level of precision. The method outlined here assumes that the same sample size (number of target mines) will be chosen for each detection system. For a more general method, see Fleiss [1981].

As a first step, the experimenter must specify four different probabilities. Using the terminology of section 2.2, the first two of these probabilities are α (the level at which the test is to be conducted) and $1-\beta$ (the desired power of the test). The remaining two probabilities are P_{A_0} and P_{B_0}, nominal detection probabilities for systems A and B such that the test will have a chance of at least $1 - \beta$ of giving a significant result when the true value of P_A equals P_{A_0} and the

true value of P_B equals P_{B_0}. Given values of α, $1 - \beta$, P_{A_0}, and P_{B_0}, a sufficient sample size for each system can be computed.

A concrete example serves to clarify the situation. Suppose that the experimenter wishes to conduct hypothesis tests at level $\alpha = 0.05$, meaning that there is a 5% chance of rejecting H_0 when the two systems do not actually differ. Suppose further that the experimenter wants to have at least an 80% chance of obtaining a significant result (rejecting H_0) if the true detection probability for system A is 0.90 and the true detection probability for system B is 0.60. Then the probabilities α, $1-\beta$, P_{A_0}, and P_{B_0} are set to 0.05, 0.80, 0.90, and 0.60, respectively.

Once the four probabilities have been specified, the required sample size can be computed. Let $z_{1-\alpha/2}$ and z_β represent the quantiles of the normal distribution corresponding to probabilities $1-\alpha/2$ and β, respectively. Compute the initial quantity:

$$n' = \left[\frac{z_{1-\alpha/2}\sqrt{(P_{A_0} + P_{B_0})(2 - P_{A_0} - P_{B_0})/2} - z_\beta \sqrt{P_{A_0}(1-P_{A_0}) + P_{B_0}(1-P_{B_0})}}{P_{B_0} - P_{A_0}} \right]^2 \qquad (23)$$

The required minimum sample size is:

$$n = n' + \frac{2}{|P_{A_0} - P_{B_0}|} \qquad (24)$$

A table in Fleiss [1981] gives values of (24) for many choices of level, power, and nominal detection probabilities. Each of the detection systems should be presented with at least n mines for the test to have the desired power.

For the probabilities specified above, (23) and (24) give $n' = 31.50$ and $n = 38.16$, implying that 39 or more mines are needed to ensure at least an 80% chance of detecting a difference when the two systems have detection probabilities of 0.90 and 0.60. The required sample size increases rapidly when the detection capabilities of the two systems converge.

Keeping α = 0.05, 1–β = 0.80 and P_{A_0} = 0.90, (24) gives sample sizes of 72 for P_{B_0} = 0.70, 219 for P_{B_0} = 0.80, and 726 for P_{B_0} = 0.85. This result is in agreement with the intuitive concept that small differences are more difficult to detect than large differences.

5 – DISCUSSION AND SUMMARY

The statistical methods outlined in this report can assist both in designing experiments to study the capabilities of different mine detection systems, and in analyzing the data collected in such experiments. In the interest of brevity, some pertinent considerations have been neglected. A few of these merit brief attention at this time.

When more than two different mine detection systems are included in the experiment, the issue of multiple comparisons becomes relevant. Consider 95% confidence intervals for a difference in mine detection probabilities. In the case where just two detection systems are under test, a single confidence interval can be constructed that will have a 95% probability of containing the true difference for these two systems. However, if some number $k > 2$ of systems are tested, then $k(k–1)/2$ pairs can be compared, each using a different 95% confidence interval. Any one of these intervals has a 95% a priori probability of containing the true difference for the two systems in the pair. However, the probability that all $k(k–1)/2$ intervals will *simultaneously* include their true parameter values is somewhat less. The text by Box, Hunter, and Hunter [1978] describes several statistical techniques developed to address the problem of multiple comparisons. Often, no special action is needed as long as the investigator is aware of the issue when interpreting the data.

Another consideration involves controlled experimental factors that are expected to affect system performance. In a typical experiment, detection systems are tested against various different target types (e.g., anti-tank and anti-personnel mines) under different scenarios (e.g., dry and moist soil). The computations and analyses outlined in this report are to be conducted separately for each combination of factors tested. In addition, the statistical method known as

the analysis of variance (ANOVA) can be used to determine which factors (mine size, mine type, burial depth, soil moisture, etc.) impact performance [Box, Hunter, and Hunter, 1978].

For many mine detection systems, the primary output is not a binary ("mine/no mine") decision, but rather a continuous variable, expressed as a function of location (or time), that is thresholded to determine where (or when) detections have occurred. In this report, only the binary results have been considered. It is assumed that threshold levels are chosen by the operators of the various systems to optimize performance. If thresholded data were not available, one logical choice would be to set a threshold for each system that corresponds to a fixed false alarm rate, and then to comparatively assess PD across systems at that chosen level of false alarms.

Any useful analysis of mine detection system performance must consider false alarm rates in addition to detection probabilities. The use of receiver-operator characteristic (ROC) curves is a standard way to display PD performance at multiple levels of FAR [Andrews, George, and Altshuler, 1997; Poor, 1988]. Additional statistical methods relevant to the analysis of false alarm data are discussed in the companion to this report [Simonson, 1998].

6 - REFERENCES

Andrews, A.M., George, V., and Altshuler, T.W. (1997). Quantifying performance of mine detectors with fewer than 10,000 targets. *SPIE Proceedings*, vol. 3079, 273-280.

Bickel, P.J., and Doksum, K.A. (1977). Mathematical Statistics. Oakland, CA: Holden-Day.

Beal, S.L. (1987). Asymptotic confidence intervals for the difference between 2 binomial parameters for use with small samples. *Biometrics*, vol. 43, 941-950.

Blyth, C.R., and Still, H.A. (1983). Binomial confidence intervals. *JASA*, vol. 78, 108-116.

Box, G.E.P, Hunter, W.G., and Hunter, J.S. (1978). Statistics for Experimenters. New York: John Wiley & Sons.

Cox, D.R., and Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

D'Agostino, R.B., Chase, W., and Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial proportions. *American Statistician*, vol. 42, 198-202.

Feller, W. (1950). An Introduction to Probability Theory and Its Applications, Volume I, Third Edition. New York: John Wiley & Sons.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, Second Edition. New York: John Wiley & Sons.

Hahn, G.J. and Shapiro, S.S. (1967). Statistical Methods in Engineering. New York: John Wiley & Sons.

Johnson, N.L., Kotz, S., and Kemp, A.W. (1992). Univariate Discrete Distributions, Second Edition. New York: John Wiley & Sons.

Koopmans, L.H. (1987). Introduction to Contemporary Statistical Methods. Boston: Duxbury Press.

Larsen, R.J., and Marx, M.L. (1981). An Introduction to Mathematical Statistics and Its Applications. Englewood Cliffs, NJ: Prentice-Hall.

Lehmann, E.L. (1975). Nonparametrics. Oakland, CA: Holden-Day.

Mantel, N. (1988). Exact limits on the ratio or difference of 2 independent binomial proportions. *Biometrics*, vol. 44, 623.

Peksun, P.H. (1993). A new confidence interval method based on the normal approximation for the difference of two binomial probabilities. *JASA*, vol. 88, 656-661.

Poor, H.V. (1988). An Introduction to Signal Detection and Estimation. New York: Springer-Verlag.

Santner, T.J., and Snell, M.K. (1980). Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. *JASA*, vol. 75, 386-394.

StatXact-3 for Windows, (1995). Software for Exact Nonparametric Inference. Cambridge, MA: Cytel Software Corporation.

Silvey, S.D. (1975). Statistical Inference. London: Chapman and Hall.

Simonson, K.M. (1998). Statistical considerations in designing tests of mine detection systems: II – Measures related to the false alarm rate, SAND98-1769/2. Sandia National Laboratories, Albuquerque, NM, August 1998.

Wallenstein, S. (1997). A non-iterative accurate asymptotic confidence interval for the difference between two proportions. *Statistics in Medicine*, vol. 16, 1329-1336.

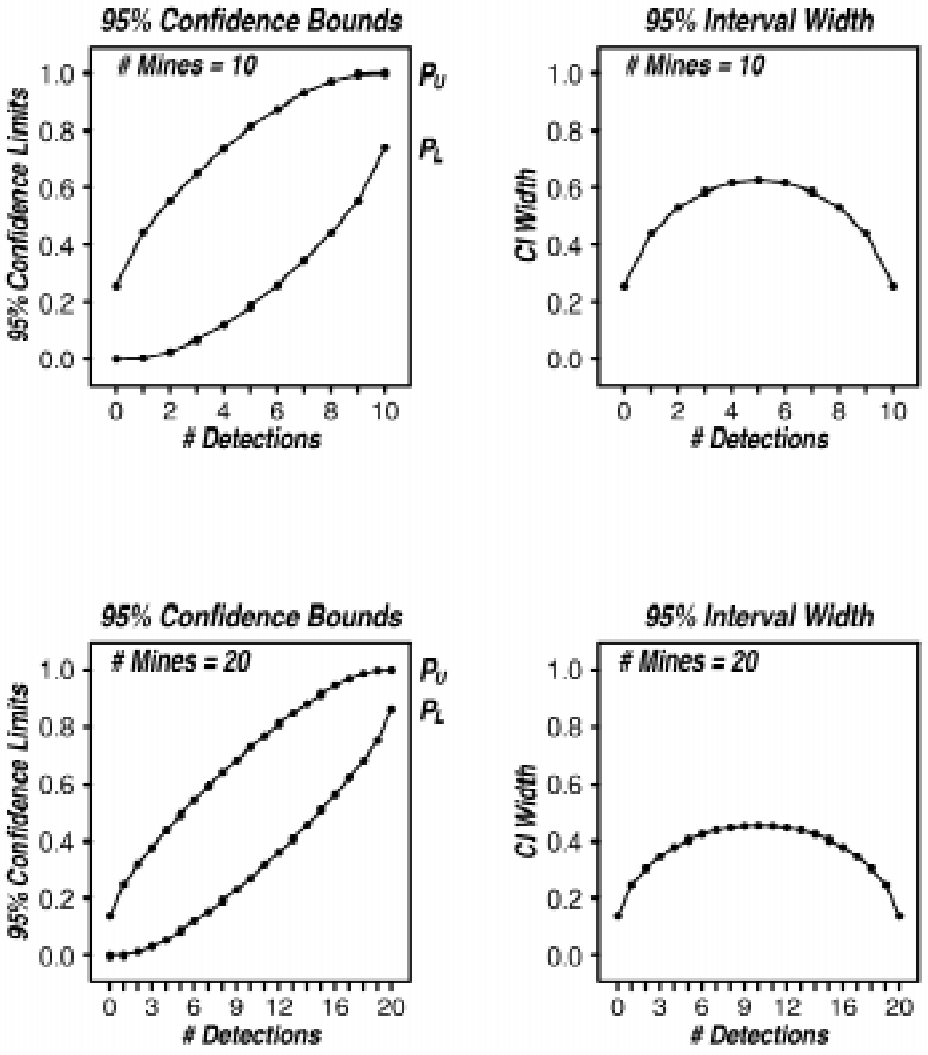

Figure 1 – Small sample confidence intervals for a single binomial parameter. Bounds and widths are shown for 95% confidence intervals, for n=10 and n=20 mines placed. All of the values plotted were computed using the small sample method of equations (4) through (9).

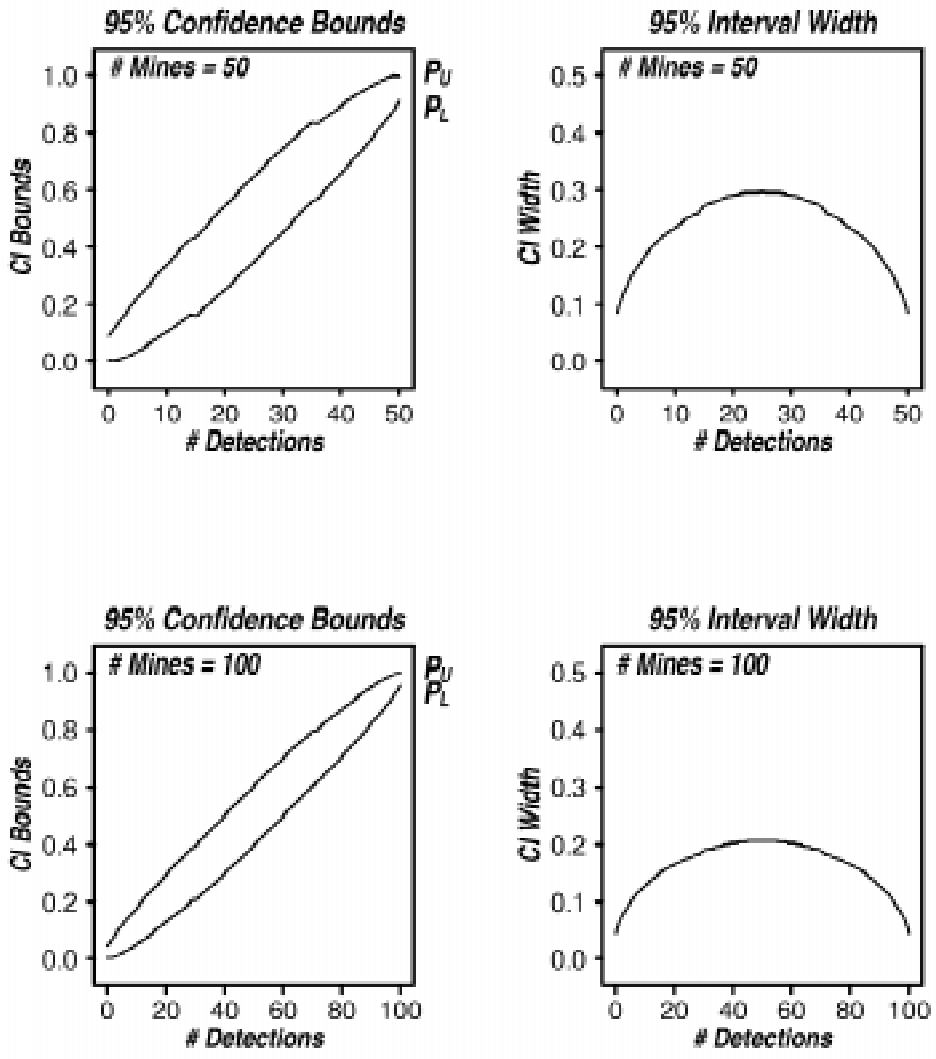

Figure 2 - Large sample confidence intervals for a single binomial parameter. Bounds and widths are shown for 95% confidence intervals, for n=50 and n=100 mines placed. All of the values plotted were computed using the large sample method of equations (10) through (13). The small jumps seen in the bounds for n=50 at $\hat{p}$ = 0.30 and $\hat{p}$ = 0.70 are due to transitions between the use of (10) and (11) and the use of (12) and (13).

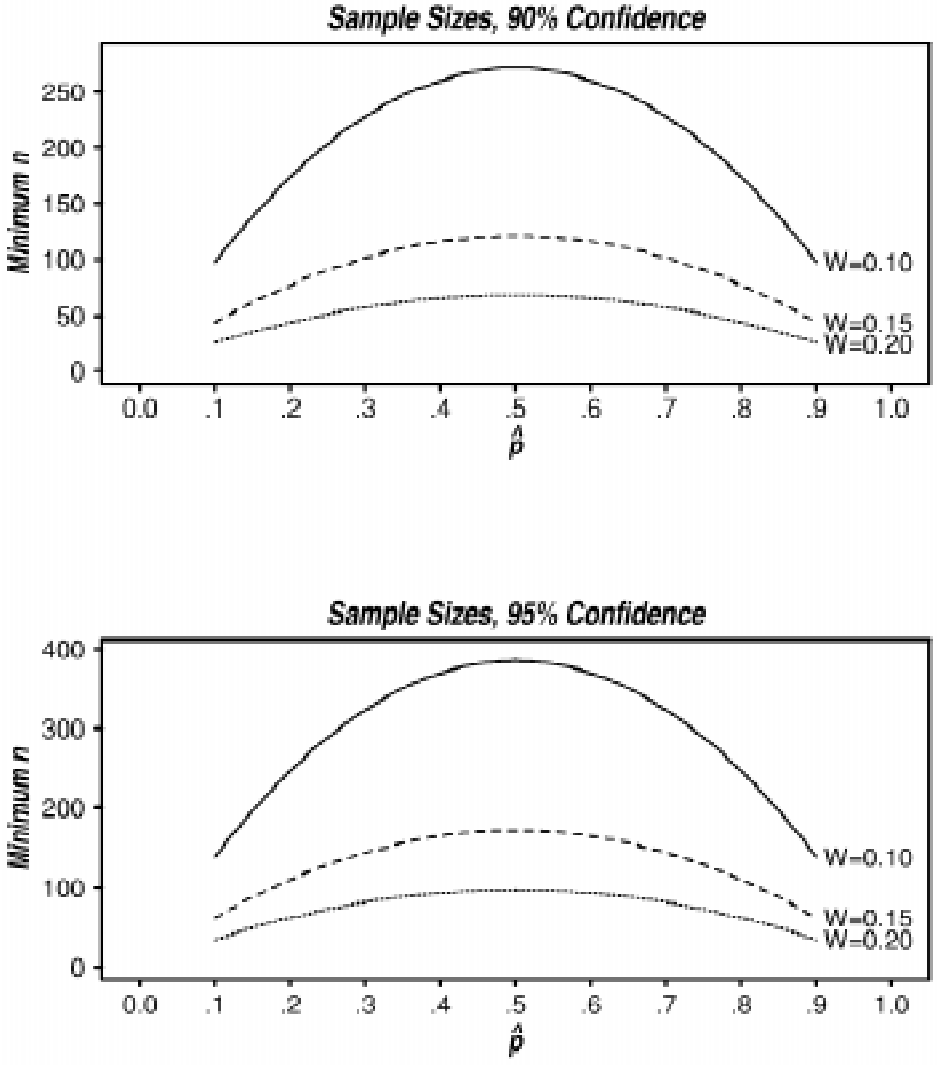

Figure 3 - Sample size requirements for the estimation of a single binomial parameter. Values plotted on the upper panel represent the number of samples needed to obtain a 90% confidence interval with width equal to 0.10, 0.15, or 0.20. The same information for 95% confidence intervals is shown in the lower panel.

DISTRIBUTION:

1	MS0844	Larry Hostetler, 2523
35	0844	Katherine Simonson, 2523
1	9018	Central Technical Files, 8940-2
2	0899	Technical Library, 4916
2	0129	Review & Approval Desk, 12690 For DOE/OSTI